International Journal of Electrical and Computer Engineering (IJECE)
Vol. 8, No. 5, October 2018, pp. 2804~2811
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i5.pp2804-2811

Forensic and Automatic Speaker Recognition System

Satyanand Singh
School of Electrical and Electronics Engineering, Fiji National University, Fiji Island

Article Info

Article history:

Received Nov 3, 2017
Revised Jan 19, 2018
Accepted Sep 29, 2018

Keyword:

Automatic Speaker Recognition
Gaussian Mixing Model
Normalization
Universal Background Model
Within-Class Covariance

ABSTRACT

Current Automatic Speaker Recognition (ASR) System has emerged as an important medium of confirmation of identity in many businesses, e-commerce applications, forensics and law enforcement as well. Specialists trained in criminological recognition can play out this undertaking far superior by looking at an arrangement of acoustic, prosodic, and semantic attributes which has been referred to as structured listening. An algorithm-based system has been developed in the recognition of forensic speakers by physics scientists and forensic linguists to reduce the probability of a contextual bias or pre-centric understanding of a reference model with the validity of an unknown audio sample and any suspicious individual. Many researchers are continuing to develop automatic algorithms in signal processing and machine learning so that improving performance can effectively introduce the speaker's identity, where the automatic system performs equally with the human audience. In this paper, I examine the literature about the identification of speakers by machines and humans, emphasizing the key technical speaker pattern emerging for the automatic technology in the last decade. I focus on many aspects of automatic speaker recognition (ASR) systems, including speaker-specific features, speaker models, standard assessment data sets, and performance metrics.

Corresponding Author:

Satyanand Singh,
School of Electrical and Electronics Engineering, CEST,
Fiji National University, Fiji Island.
Email: satyanand.singh@fnu.ac.fj

1. INTRODUCTION

Speaker recognition and verification have achieved visibility and significance in society as speech technology, audio content, and e-commerce continues to expand. There is an increasing need to search audio content and start research based on the speaker identity is increasing interest to a young scientist. Imagining the future is not difficult when a system will expose our identity not only the sense of the intelligent, sympathetic and fully functional personal assistants, which we will say, but by our voice, we recognize more track-able or other recognizable symptoms.

This is the additional basic information that we can not recognize the voice of a person once heard and at the same time, it is difficult to identify the voice of a known person on the telephone. In view of these thoughts, a native person may ponder what precisely makes speaker recognition such a difficult task and why is it a point of such thorough research. From the above discussion, we can say that the identity of the speaker can be completed in three steps. Any individual can easily recognize the familiar sounds of a person without any conscious training. These methods of recognition can be called as "Native Speaker Recognition". In the forensic identification, a voice sample of a person from telephone calls database is often compared with potential suspects. In these cases, there are trained listeners in order to provide a decision. We will categorize this method as forensic speaker recognition. In this computer-based world, we have an automatic speaker recognition system, where an electronic machine is used to complete a speech analysis and automated decision-making. Forensic and ASR research communities have developed several methods for at least seven

decades independently. In contrast, native recognition is the natural ability of human beings which is always very effective and accurate. Recent research on brain imaging has shown many details that how a human being does cognitive-based speakers recognition, which can motivate new directions for both automated and forensic system [1, 2]. In this review paper, I present a as on date literature review of ASR systems, especially in the last seven decades, providing the reader with an attitude of how the forensic by the human speaker, especially the expert, and the native audience recognize. Its main purpose is to discuss three said sections of speaker recognition, which are important similarities and differences between them. I insist on how automatic speaker recognition system has been developed on more current approaches over time. In noise masking, many speech processing techniques, such as Mel scale filter bank for feature extraction and concepts, inspired by the human hearing system. Also, there are parallels between forensic voice experts and methods used by automated systems, however, in many cases, research communities are different. I believe that it required to include in this review, the perspective of the concept of speech by humans, including highlights of both strengths and weaknesses in speaker recognition system compared to machines, it will help readers to see and perhaps inspire new research in the field of the man-machine interface.

In the first place, to consider the general research domain, it is valuable to elucidate what is enveloped by the term speaker recognition, which comprises of two task undertakings: verification and recognition. In speaker recognition, the undertaking is to distinguish an obscure speaker from an arrangement of known speakers. As it were, the objective is to find the speaker who sounds nearest to the speech coming from an obscure speaker inside a speech database. At the point when all speakers inside a given set are known as a closed set situation. On the other hand, if the potential information from outside the predefined known speaker gathering, this turns into an open-set situation, and, hence, a world model or universal background model (UBM) [3] is required. This situation is called open-set speaker recognition.

2. THE MAIN CHALLENGES IN AUTOMATIC SPEAKER RECOGNITION SYSTEM IN PRESENT SCENARIO

For example, like other biometric systems, iris, finger, face, and hand [4], the human voice is also a demonstration of the biometric system. The identity of the narrator is naturally embedded and specifically how a dialect is spoken for a person, not necessarily what is being said. This increases the possibility of speech signals with the degree of variability.

If a person does not say the same word exactly the same way then it is called inter-speaker variability [4, 5]. In addition, various electronic devices used in recording and transmission methods usually increase the system complexity. A person may find it hard to identify a person's voice through a mobile, or when a person suffers from cold and he/she is not healthy or he/she is performing another work in a stressed situation. The source of variability of speakers can be broadly classified into three categories: (i) Technology-based, (ii) Speaker-based, and (iii) Conversations based.

2.1. Challenge and Opportunity in Speaker Recognition

Technology is more to focus the initial efforts in speaker recognition, which includes telecommunications sector, where the communications channel and telephone handset variation was the main concern. Smartphone dominate the telecom industry, the variety of telephony landscape has expanded significantly. Speaker option available with all smartphone makes the user interact at a distance from the microphone, and this initiated a broad range of variability in the channel. The performance of speaker recognition system depends on intersession variability as well as the inherent changes present within human utterances recorded at the different session. However, the speaker recognition efficiency seems to be independent of time of voice samples collected for training and testing purpose [6, 7].

Most of the forensic speaker recognition uses in different legal scenarios are not very complicated. When adequate voice samples are available from the criminal, then methodical study can be done to extract the speaker specific properties, which are also called speaker specific feature parameter from voice data, and can be compared between the samples. In automatic speaker recognition system speaker-specific features were extracted from the speech signal and mathematically modeled to perform a meaningful comparison.

2.2. Individual Characterization Based on Speaker Specific Features

Every individual in the world has certain character traits in his/her speech that is unique. Speaking characteristics of an individual cannot be so different from the other, but mainly the speaker vocal tract is unique due to the physiology and due to the learning habits of expression. Even a twin has differences in his or her voice, though according to the research he or she has the same vocal tract size [8] and acoustical properties [9], and it is difficult to separate them from conceptual/forensic perspective [10, 11]. Thus, whether the speaker is identified by humans or machines, unambiguous aspects of some measurable and

predefined speaker-specific features should be considered for meaningful comparison in speech. In general, we prefer these characterizing aspects as feature parameters in human speech signal.

No one can expect that a unique speech signal of a person should be unique features, but it is not always true. Let us consider two different speakers with equal speaking rate with s a suitable feature with differ pitch. It is complicated by the intra-variability and degradations discussed earlier, this is why many feature parameters are important. Nolan has reported in his article ideal speaker specific feature parameter must have these properties [12]: easy to extract and process, robust, high frequency of occurrence, highly resistive to attempted disguise or mimicry. Speaker-specific feature parameters can be classified into short-term versus long-term, linguistic versus nonlinguistic, and auditory versus acoustic features. There are strengths and weaknesses of auditory and acoustic features. Two samples of the speech signal may sound very similar, but acoustic parameters differ greatly [13].

3. FORENSIC SPEAKER RECOGNITION

Identification of forensic speakers needs to recognize the problem occurs when you leave your voice as criminal evidence, a telephone recording or an audible speech by ear witness. Through the recognition technology, forensic speakers were discussed with speech waves that 1926 [14]. Later, spectrographic was developed representing speech at AT & T Bell Laboratories during World War II. Much later in 1970, when it came to be known as a voice print [15]. As the name shows, voice print has also been presented with fingerprints and very high expectations.

Later, the reliability of voice printing for speech recognition to its operating system, the formal process, examined and fully supported [16, 17] which "is an idea that has gone wrong," said [17]. Today, most researchers believe that it is better controversial. Voiceprint a chronological history is found in [18] and an overview of the discussion are found in forensic speaker recognition [19] here I present an overview of current trends [4]. Today, forensic recognition is performed by the expert generally phoneticians which are typically in the linguistic and statistical background.

3.1. Different Approaches to Forensic Speaker Identification

The described methods are done by human experts in whole or in part. While they are also considered for the forensic speaker recognition by the complete automated approach, we discuss the automatic identification of speakers in later sections. The auditory phonetician's approach is based on human auditory system and based on their experience they produce a detailed transcript of the test samples. Forensic experts try to hear specimen sampling and detect any presence of unusual sounds, specific or noteworthy [20]. Expert experience is evidently the main aspect in scarce or typical decision-making. The above discussed auditory functions are used in this approach.

As long as it is combined with other methods of hearing approach, it is completely subjective. Althoughthe Likelihood Ratio (LR) can be used to express results, forensic expert generally do not use the auditory approach. Instead, on the basis of their comparison of auditory actions, they present a statement of evidence in the court. The auditory spectrogram approach is derived from the voice known in the same word or phrase and their spectrograms are visually analyzed. After the debate over voiceprint, the spectrographic technique developed. If this explains, then forensic experts did not have the spectrographs separating variability by intraspeaker and interspeaker by a normal view assessment. So they have developed different protocols to analyze the aspects of pre-determined spectrographs that require the forensic examiner.

3.2. Speaker Recognition by Human

The skill to distinguish people by listening voice is a God's gifted characteristics. It mentioned in the "Mahabharata" which some historians say was written in 400 BC that when Abhimanyu was in his mother's womb,Sri Krishna used to walk around Shubhadra. To humour her, Krishna used to relate many of his adventures to the pregnant Subhadra. On this excursion, Krishna described his experience with the Chakra-Vyu technique and how it could be inserted step by step in various circles could be penetrated. However, it seems that Subhadra did not find this interesting topic and fell asleep early. However, someone else was interested in the description of Shri Krishna so far Abhimanyu was not born. We use spectral features, including language, prosody, and lyrical style, to identify a number of different aspects of the human voice, to identify a person. Even without a conscious effort, do not forget to remember these features. There are various aspects in which the inexperienced listener is currently known about how to make specific speaker recognition based on these aspects (i) Voice segment identification (ii) Recognition and discrimination (iii) Language familiarity (iv) Abstract representation of speech.

3.3. State-of-the-Art Automatic Speaker Recognition System

ASR is a mathematical algorithm based computer system designed to recognise the voice of a speaker operated independently with minimum human intervention. The ASR system admin can adjust algorithm parameters, but to compare between speech segments, all users have to provide speech signal to the ASR system. In this paper, I concentrate attention on the text-independent ASR system and the speaker verification. As mentioned earlier, humans are good in differentiating voiced and non-voiced signal that is the important part in auditory forensic speaker recognition. Obviously, in ASR it is desirable that the speaker-specific feature can only be extracted from the voiced speech signal by voice activity detection (VAD) [21, 22]. Detection and feature extraction from speech segment is important when considering the condition of excessive noise/degraded speech signal. Recently used VAD algorithm is explained in [21] although more accurate unsupervised solution Speech Activity Detected (SAD) has emerged as successful in various ASR applications in diverse audio condition [23].

Short-term speaker specific feature in ASR application shows the parameters extracted from the short segment of speech signal within 20-25 ms. In ASR application the most popular short-term acoustic features reported are the Mel-frequency cepstral coefficients (MFCCs) [24] and linear predictive coding (LPC) based features [25]. Steps involved in to obtain MFCC feature from speech signal are (i) Divide speech signal into short overlapping form (25 ms). (ii) Multiplication of these segments with Hamming and Hanning window function to get Fourier power spectrum (iii) Apply logarithm of the spectrum (iv) Apply nonlinear Mel-space filter-bank to obtain spectral energy in each channel (24 channel filter bank) (v) Apply discrete cosine transform (DCT) to obtain MFCC. As previously indicated, the specific speaker feature is the desirable qualities of the acoustic feature are robustness to degradation. The features normalization is one of the desirable characteristics of an ideal feature parameter [26].

4. MODELING OF STATE-OF-THE-ART ASR SYSTEM

Converting audio segments into the functional parameter, after that modeling process started in ASR. In ASR modeling is a process flow to categories all speakers based on their characteristics. The model should also provide its meaning for comparison with unfamiliar speaker utterances. ASR modeling is called as robust when its speaker specific feature characterization process is not significantly affected by unwanted maladies, although these features are ideal if such features can be designed in such a way that interspeaker discrimination is maximum, then no intraspeaker variation exists and simple modeling methods can be sufficient. In short form, the non-ideal properties of the speaker specific feature extraction phase require different compensation techniques during the ASR modeling phase so that the effect of the disturbance variation present in the speech signal can be reduced during the testing of the speaker recognition process. Most of the ASR modeling techniques do different mathematical hypotheses about the speaker-specific features. If assumed properties are not met from the speech data, then we are basically presenting flaws even during the ASR modeling phase.

The normalization of speaker-specific features can reduce these problems to some extent, but not completely. As a result, mathematical models are compelled to adopt the characteristics and speaker recognition scores are obtained based on these models and test speech data. Thus, in this process, the properties of detecting artifacts are introduced and a family of score standardization techniques has been proposed which is proposed to complete this final stage mismatch [27]. In essence, the decline in acoustic signal affects the speaker-specific features, patterns, and scores. Therefore, it is important to improve the robustness of ASR systems in all three domains. It has been mentioned recently that speaker modeling techniques have improved and score normalization techniques are not much effective [28, 29].

4.1. ASR System Based on Gaussian Mixture Model (GMM)

When there is no prior knowledge of speech content in text-independent speaker recognition tasks, it has been found that GMM applications are more effective for acoustic modeling to shape short-term functionality. The average behavior of this is expected short-term spectral features are more dependent on speakers than being influenced by the temporary features. Therefore, even when the test data of ASR has a different acoustic situation, then due to GMM being a potential model it may be related to better data than the more restrictive Vector Quantization(VQ) model. A GMM is a mixture of Gaussian probability density functions (PDFs), parameterized by a number of mean vectors, covariance matrices, and weights of the individual mixture components. The template is a weighted sum of individual PDFs. The density of the Gaussian mixture is the weighted sum of M component densities and it represented mathematically:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (1)$$

Where $\vec{x}$ represents D-dimension random vectors, component densities $b_i(\vec{x}), i = 1,..,M$, and mixture weight represented by p_i. Each component density is a D vector Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}\,|\sum_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)'\sum_i^{-1}(\vec{x}-\vec{\mu}_i)\right\} \quad (2)$$

$\vec{\mu}_i$ represents mean vector, $\sum_i$ represents covariance matrix. The complete density of the Gaussian mixture is parameterized by the mean vector, covariance matrix and mixture components of all density. These parameters are represented collectively by signaling

$$\lambda = \{p_i, \vec{\mu}_i, \sum_i\}\ i = 1,..,M \quad (3)$$

For ASR system, each speaker is represented by one by the GMM and is referred to by his/her model λ. The size of GMM may vary depending on the choice of covariance matrix. The GMM model can be evaluated using the probability of a vector attribute in eqn. (1).

4.2. Support Vector Machines (SVMs)

An SVM is a binary classifier that makes its decisions by constructing a linear decision boundary or hyperplane that optimally separates the two classes. Depending on its position in relation to Hyperplane, the model can be used to predict the class of unknown observation. Let us consider training vector and labels as (x_n, y_n) , $x_n \in \Re^d$, $y_n \in \{-1, +1\}$, $n \in \{1, \ldots T\}$ the optimal hyperplane is chosen according to the maximum margin criterion then target of SVM can be learn the function $f: \Re^d \to \Re$ so that the class labels of any unknown vector x can be expected as $I(x) = \text{sign}(f(x))$.

For linearly separable data labeled [5, 30], hyperplain H can be obtained from $x^T x + b = 0$, which separates the two class of data, so that $y_n(w^t x_n + b) \geq 1$, $n \ldots . T$. An optimal linear divider H provides maximum margins between classes, i.e. the distance between H and the training of two different sections is highest in the data estimates. The maximum margin is found in the form of $\frac{2}{\|w\|}$ and data points x_n for which $y_n(w^t x_n + b) \geq 1$ that the margin is known as super vectors. When ASR training data is not linearly separable, then speaker specific features can be mapped to a higher dimensional space, in which kernel functions are linearly divided.

4.3. Factor Analysis (FA) of the GMM Supervectors

The purpose of the FA is to describe variability in high dimensional observable data vector using less number of unobservable/hidden variables. For ASR application, the idea of explaining peaker's and channel-dependent variability in the GMM supervector space, FA has been used in [31]. Many forms of FA methods have been employed since, which ultimately brought the current state of the art i-vector approach. In a linear distortion model, a speaker-dependent GMM supervisor m_s is generally considered as four component which are linear in nature.

$$m_{s,h} = m_0 + m_{spk} + m_{ch} + m_{res} \quad (4)$$

Where m_0 is speaker channel environment-independent component, m_{spk} is speaker dependant component, m_{ch} is channel environment dependant component and m_{res} is residual. The joint FA (JFA) model is prepared in conjunction with eigenvoice and eigenchannel, which is achieved with a MAP optimization for a model. The sub-spaces are aligned by V and U matrix, as the first model recommends for an informal choice of speakers s and sessions h, mean supervector of GMM can be represented by

$$m_{s,h} = m_0 + U_{Xh} + V_{ys} + D_{Zs,h} \quad (5)$$

So now this is the only model, which we are considering all the four components of linear distortion model we discussed earlier. In fact, JFA has been shown to overcome other current method.

4.4. i-Vector Approach

In an effort to unify the strength of these two methods, modern ASR systems attempted to utilize JFA as a speaker specific feature extractor by Dehak et al. [32] for SVM. In the initial effort speaker factors estimation JFA were used as speaker specific feature for the SVM classifiers. Keeping in mind that even channel factors have information of speakers and the channel has been added to a single space, called total

variability space [33]. The FA model that depends on, speaker and session is represented by a GMM supervisor as

$$m_{s,h} = m_0 + T_{Ws,h} \quad (6)$$

$T_{Ws,h}$ is called total factor. Like all the FA methods described above, hidden variables are not overlooked, but their posterior expectation can be estimated. The total factor estimate, which can be used as features in the next stage of the classifier named as i-vectors.

4.5. Linear Discriminant Analysis (LDA) approach

LDA is a commonly employed technique in statistical pattern recognition that aims at finding linear combinations of feature coefficients to facilitate discrimination of multiple classes. It finds orthogonal orientation in place of most effective functions in class discrimination. By introducing the original features in these guidelines, the accuracy of classification improves. Let us indicate set of all development utterances by D, utterance features indicated by $w_{s,i}$, these features obtained from the ith utterance of speaker s, the total number of utterances belonging to s is indicated by n_s and total number of speakers in D is indicated by S. Class covariance matrices between S_b and within S_w are given by

$$S_b = \frac{1}{S}\sum_{s=1}^{S}(\bar{w}_s - \bar{w})(\bar{w}_s - \bar{w})^T \quad (7)$$

$$S_w = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}\left(w_{s,i} - \bar{w}_s\right)\left(w_{s,i} - \bar{w}_s\right)^T \quad (8)$$

Where the speaker dependant mean vector is given by $\bar{w}_s = 1/n_s \sum_{i=1}^{n_s} w_{s,i}$ and speaker independent mean vector is given by $\bar{w} = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s} w_{s,i}$ respectively. The LDA optimization is therefore to maximize between class variance, whereas reducing within the class variance. The exact estimation can be obtain from this optimization by solving generalized eigenvalue problem:

$$S_{bV} = \wedge S_w v \quad (9)$$

The diagonal matrix containing of eignvector is indicated by ∧. If the matrix S_w in eqn. (8) is invertible then the solution can be easily found by $S_w^{-1}S_b$. A_{LDA} matrix of dimension $R \times k$ is as follows

$$A_{LDA} = [v_1 \dots \dots \dots v_k] \quad (10)$$

k eigenvectors $v_1 \dots \dots \dots v_k$ obtained by solving eqn. (9). Thus, the LDA change of the utterance feature w is obtained in this way

$$\Phi_{LDA}(w) = A_{LDA}^T w \quad (11)$$

4.6. Nuisance Attribute Projection (NAP)

The application of NAP algorithm in ASR reported in [34]. In NAP technique the speaker specific feature space is replaced by complementary channel space using an orthogonal projection, which depends only on the speaker. The projection matrix of size $d \times d$ is calculated using covariance matrix of co-rank $k < d$ as $P = I - u_{[k]}u_{[k]}^T$. The low rank rectangular matrix $u_{[k]}$ whose column is k principal eigenvectors of the within-class covariance matrix S_w in eqn. (8). The NAP is performed on w as $\Phi_{NAP}(w) = Pw$.

4.7. Within-Class Covariance Normalization (WCCN)

The main goal of WCCN normalization to improve the robustness of the SVM-based ASR framework [35] using a consistent opposite decision approach. The aim of the WCCN launch is to reduce false alarm rates and miss-errors rates during SVM training. Covariance matrix within-class S_w is calculated using eqn. (8) and projection on WCCN is performed as $\Phi_{WCCN}(w) = A_{WCCN}^T w$. With the help of Cholesky factorization of S_w^{-1} A_{WCCN} is computed as $S_w^{-1} = A_{WCCN}A_{WCCN}^T$. Unlike LDA and NAP, the projection of WCCN easily converses the feature space.

5. ASR PERFORMANCE EVALUATION IN STANDARD SPEECH DATA SETS AND TYPES OF ERROR

Performance evaluation of ASR system is one of the main aspects of the research cycle. It is strongly dependent on the variability of the voice signal, noise and distortion in the communication channel. Recognition has to face many problems: unrestricted input speech, non-co-operative speaker and uncontrolled environmental norms. There are two types of errors may occur in such decision making processes in ASR system (i) false rejection (in other words non-detection), that is, the system disapproves a genuine identity claim of a speaker under scrutiny and (ii) false acceptance (in other words false alarm), that is, the system approves the identity claim of an impostor.

These errors are quantified as performance measures of a security system. They are (i) False Rejection Rate (FRR), which indicates the percentage of incorrectly rejected clients and (ii) False Acceptance Rate (FAR). In a real life situation, a biometric security system, which is usually imperfect, the characteristic curves of FRR and FAR intersect at a certain point called 'Equal Error Rate (EER)'. If one fixes a very low threshold value, then the system would exhibit very low FRR and very high FAR and accept all identity claims. Alternatively, if one fixes a very high threshold value, then the system would exhibit very high FRR and very low FAR and reject all identity claims. In this context, one could plot a curve called 'Receiver Operating Characteristic (ROC)', which involves FRR and FAR. ROC curve is a graphical indication of the system performance.

As mentioned above, EER does not distinguish between two types of errors which are sometimes unrealistic performance evaluation of ASR. Therefore, the detection cost function (DCF) introduces the numerical/penalty cost for two types of errors. The priori probability of encountering a target speaker provides priority and DCF is calculated as the decision threshold value as $\mathrm{DCF}(\tau) = \mathrm{C_{MISS}}\mathrm{P}(\tau)\mathrm{P_{target}} + \mathrm{C_{FA}}\mathrm{P_{FA}}(\tau)(1 - \mathrm{P_{target}})$. Where Cost of a miss/FR error is indicated by $\mathrm{C_{MISS}}$, Cost of an FA error is indicated by $\mathrm{C_{FA}}$, Prior probability of target speaker is indicated by $\mathrm{P_{target}}$, Probability of $(\mathrm{MISS}|\mathrm{Target},\ \mathrm{Threshold} = \tau)$ is indicated by $\mathrm{P_{miss}}(\tau)$ and Probability of $(\mathrm{FA}|\mathrm{nontarger},\ \mathrm{Threshold} = \tau)$ is indicated by $\mathrm{P_{FA}}(\tau)$.

The above three quantities in NIST SRE 2008 $\mathrm{C_{MISS}} = 10$, $\mathrm{C_{FA}} = 1$ and $\mathrm{P_{target}} = 0.01$ are predefined. In general, the goal of the ASR system designer is to find the optimum threshold value which reduces the DCF. Now, the prior value $\mathrm{P_{target}} = 0.01$ indicates that ASR system will be detected after every 100 attempts to check the speaker. When the speaker recognition performance is evaluated in different operational points, then the error detection curve (DET) is usually used. DET curve is a FAR error plot compared to FRR/miss. When the performance of the ASR system improves, the curve moves toward origin. The DET curve nearest to origin represents a better ASR system.

6. CONCLUSION

There is still a lot of work to fully understand the way to decide on the content of human brain speech and speakers. However, what we know, it can be said that the ASR system should focus on improving performance, more on high-level speaker-specific features. Human beings are effective in the identification of unique speakers; they know it very well, while ASR systems can only learn a specific section if a measurable function parameter can be defined correctly. A large number of automated systems audio is better in researching and possibly, more effective to reduce the likelihood of those audio samples being speakers matches; while humans are better to compare a smaller subgroup and do not match the microphone or channel more easily. It can be useful to check exactly what the "know" of a speaker means from a perspective of a useful system. The discovery of alternative compact speaker representations and audio segments that emphasize relevant identification parameters, while eliminating nuisance components will always be a continuous challenge for state-of-the-art ASR system developers.

REFERENCES

[1] Syeiva Nurul Desylvia *et al*, "Modeling Text Independent Speaker Identification with Vector Quantization," *TELKOMNIKA*, vol.15(1), 2017, pp. 322-327.
[2] E. Formisano, *et al*, " 'Who' is saying 'what'? Brainbased decoding of human voice and speech," *Science*, vol. 322, 2008, pp. 970–973.
[3] D. A. Reynolds, *et al*, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Proces*, vol. 10(1), 2000, pp. 19–41.
[4] John H.L. Hansen and Taufiq Hasan, "Speaker Recognition by Machines and Humans." *IEEE Signal Process. Mag.*, 2015, pp. 74-99.
[5] Amali Mary Bastina , N. Rama, "Biometric Identification and Authentication Providence using Fingerprint for Cloud Data Access," *International Journal of Electrical and Computer Engineering* , vol. 7(1), 2017, pp. 408-416.

[6] A. D. Lawson, *et al*, "The multi-session audio research project (MARP) corpus: Goals, design and initial findings," *in Proc. Interspeech, Brighton, U.K.*, 2009, pp. 1811–1814.
[7] L. A. Ramig and R. L. Ringel, "Effects of physiological aging on selected acoustic characteristics of voice," *J. Speech Lang. Hearing Res*., vol. 26, 1983, pp. 22–30.
[8] F. Nolan and T. Oh, "Identical twins, different voices," *Int. J. Speech Lang. Law*, vol. 3(1), 1996, pp. 39–49.
[9] W. D. Van Gysel, *et al*, "Voice similarity in identical twins," *Acta Otorhinolaryngol. Belg.*, vol. 55 (1), 2001, pp. 49-55.
[10] K. M. Van Lierde, *et al*, "Genetics of vocal quality characteristics in monozygotic twins: a multiparameter approach," *J. Voice*, vol. 19(4), 2005, pp. 511-518.
[11] D. Loakes, "A forensic phonetic investigation into the speech patterns of identical and non-identical twins," *Int. J. Speech Lang. Law*, vol. 15(1), 2008, pp. 97-100.
[12] F. Nolan, *The Phonetic Bases of Speaker Recognition*. Cambridge, U.K.: Cambridge Univ. Press, 1983.
[13] F. Nolan, "The limitations of auditory-phonetic speaker identification," *in Texte Zur Theorie Und Praxis Forensischer Linguistik, H. Kniffka, Ed. Berlin, Germany*: De Gruyter, 1990, pp. 457–479.
[14] J. H. Wigmore, "A new mode of identifying criminals," *Amer Inst. Crim. L. Criminology 165*, vol. 17(2), pp. 165-166, Aug. 1926.
[15] L. G. Kersta, "Voiceprint identification," *The Journal of the Acoustical Society of America*, vol. 34(5), 2005, pp. 725-735.
[16] Fajri Kurniawan , *et al*., "Statistical Based Audio Forensic on Identical Microphones," *International Journal of Electrical and Computer Engineering*, vol. 6(5), 2016, pp. 2211-2218.
[17] H. F. Hollien, *Forensic Voice Identification*. New York: Academic Press, 2002.
[18] L. Yount, *Forensic Science: From Fibers to Fingerprints*. New York: Chelsea House, 2007.
[19] J. P. Campbell, *et al*, "Forensic speaker recognition," *IEEE Signal Process. Mag*., vol. 26(2), 2009, pp. 95–103.
[20] G. S. Morrison, "Forensic voice comparison," *in Expert Evidence 99, 1 ed. London: Thompson Reuters*, 2010, Chap. 99, pp. 1051-1071.
[21] S. Singh, Abhay Kumar, David Raju Kolluri, "Efficient Modelling Technique based Speaker Recognition under Limited Speech Data," *International Journal of Image, Graphics and Signal Processing*. vol.8(11) 2016, pp.41-48.
[22] F. Beritelli and A. Spadaccini, "The role of voice activity detection in forensic speaker verification," *in Proc. Digital Signal Processing*, 2011, pp. 1–6.
[23] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett*., vol. 20(3), 2013, pp. 197–200.
[24] S.Singh and Dr. E.G. Rajan "MFCC VQ Based Speaker Recognition and Its Accuracy Affecting Factors" *International Journal of Computer Application*.vol 21(6), 2011, pp 1-6.
[25] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer*, vol. 87(4),1990, pp. 1738.
[26] Douglas Reynolds, *et al*, "The Super SID project: Exploiting high-level information for high-accuracy speaker recognition," *in Proc. IEEE Acoustics, Speech, and Signal Processing*, 2003, pp. 784-787.
[27] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Process*., vol. 10(1), 2000, pp. 42-54.
[28] S.Singh and Ajeet Singh "Accuracy Comparison using Different Modeling Techniques under Limited Speech Data of Speaker Recognition Systems, " *Global Journal of Science Frontier Research: F Mathematics and Decision Sciences*, vol 16(2) 2016, , pp.1-17.
[29] P. Kenny, "Bayesian speaker verification with heavy tailed priors," *in Proc. Odyssey: The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.
[30] S.V.S.Prasad, T. Satya Savithri, Iyyanki V. Murali Krishna, "Comparison of Accuracy Measures for RS Image Classification using SVM and ANN Classifiers," *International Journal of Electrical and Computer Engineering* ,vol. 7(3), 2017, pp. 1180-1187.
[31] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," *in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2004, pp. 37-40.
[32] N. Dehak, *et. al*, "Support vector machines and joint factor analysis for speaker verification," in *Proc. IEEE Int. Acoustics, Speech, and Signal Processing* , 2009, pp. 4237-4240.
[33] N. Dehak, *et. al*, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification." *in Proc. Interspeech*, 2009, pp. 1559-1562.
[34] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," *in Proc. IEEE Acoustics, Speech, and Signal Processing* , 2005, pp. 629-632.
[35] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," *in Proc. Interspeech, Pittsburgh, PA*, 2006, pp. 1471-1474.